UNIVERSITÉ DE NANTES

IAE NANTES
ÉCONOMIE & MANAGEMENT

Machine Learning with Python

Kaggle project : Titanic

Diane THIERRY

Enseignante : Camille GICHARD

Année universitaire : 2020-2021

Master 2 Économétrie et Statistiques, parcours Économétrie Appliquée

Table des matières

1 Introduction

Durant sa traversée depuis l'Angleterre vers les États-Unis en avril 1912, le fameux paquebot '**Titanic**' long de 269 mètres et large de 28 mètres sombre dans la nuit du 14 au 15 avril à la suite d'une collision latérale avec un iceberg dans l'océan Atlantique, alors qu'il devait arriver au port de New York quelques jours plus tard. Alors surnommé l'*incoulable*, le naufrage de ce navire provoque la mort de 1502 personnes soit 55% des passagers à son bord. Ainsi, malgré une taille et des technologies démesurées pour l'époque, le nombre de canots de sauvetage insuffisant et des erreurs dans la direction du navire ont provoqué le naufrage de l'insubmersible.

Outre le facteur chance qui a joué sur la survie de certains passagers, il semble que certaines personnes étaient plus susceptibles de survivre que d'autres. L'objectif de ce dossier est de répondre à la problématique issue de la compétition kaggle "Titanic : Machine Learning from Disaster", en modélisant les caractéristiques des passagers en fonction de leur survie ou non. Nous utiliserons pour cela à la fois des modèles économétriques et de machine learning et retiendrons le modèle le plus performant dans l'explication de la survie des passagers lors de ce (trop) fameux naufrage.

Le Machine Learning que nous utiliserons dans cette analyse (ML) est une technologie d'intelligence artificielle permettant l'apprentissage automatique aux ordinateurs par le biais de différents algorithmes mis en oeuvre par l'homme. L'application de ML *"spam filter"* ayant vu le jour dans les années 90 en est un bon exemple. Elle permettait de filtrer les emails en les triant en non désirés lorsque ceux-ci validaient un certain nombre de caractéristiques propres aux spams, de telle manière qu'il était rare de devoir les trier soi-même. Ainsi, le machine learning consiste à donner aux ordinateurs la capacité d'apprendre et de se perfectionner sans être explicitement programmés.

Dans une première partie nous manipulerons les bases pour les rendre exploitables et propres de manière à, dans une deuxième partie, explorer les données afin de cerner au mieux le sujet et les caractéristiques de la base. Puis, dans une troisième partie nous appliquerons les modèles et déterminerons quels ont été les facteurs les plus importants dans l'explication de la survie ou du décès des passagers.

2 Manipulations préliminaires de la base

2.1 Les données

Nous disposons pour cette analyse de 2 jeux de données :

- un jeu **train** sur lequel nous entraînerons nos modèles, il se compose de 891 observations et 12 variables dont celle que nous cherchons à expliquer.
- un jeu **test** sur lequel nous testerons nos modèles, il se compose de 418 observations et 11 variables (puisque le but est de prédire Y).

Chaque observation des jeux correspond à un passager avec pour chacun les informations suivantes :

- **Survived** : variable binaire indiquant s'il a survécu (1) ou non (0)
- **Pclass** : la classe de son voyage allant de 1 à 3 où 1 correspond à la classe la plus aisée
- **Name** : le nom du passager ainsi que son "titre" (Mr, Miss, Dr etc.)
- **Sex** : son sexe homme (male) ou femme (female)
- **Age** : son âge en année(s)
- **SibSp** : le nombre de frères et soeurs ou conjoints à bord
- **Parch** : le nombre d'enfants ou parents à bord
- **Ticket** : indicatif du numéro du ticket
- **Fare** : le prix auquel le passager a payé son billet, en livre sterling
- **Cabin** : le numéro de sa cabine avec une lettre correspondant à l'un des 7 ponts allant de A à G, puis le numéro de la cabine sur ce pont
- **Embarked** : sur l'image suivante nous voyons la trajectoire suivie par le paquebot, cette variable indique le port d'embarquement ; Southampton, Cherbourg, ou Queenstown

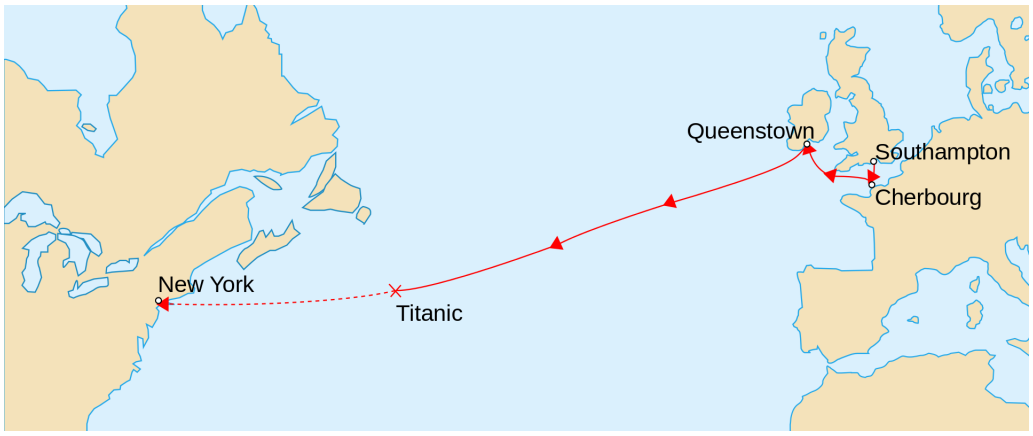


FIGURE 1 – Voyage prévu pour l'inauguration du Titanic

2.2 Manipulations des données

Avant toute analyse exploratoire des données il est important de les rendre propres pour avoir une étude viable par la suite - cela passe par la détection et la correction des valeurs manquantes et des points atypiques.

	Jeu train	Jeu test
Survie	0%	/
Classe	0%	0%
Nom	0%	0%
Sexe	0%	0%
Âge	19.87%	20.57%
Fratrie/conjoints	0%	0%
Enfants/parents	0%	0%
Ticket	0%	0%
Prix du billet	0%	0.24%
N° cabine	77.10%	78.23%
Port d'embarquement	0.22%	0%

TABLE 1 – Pourcentage de valeurs manquantes par variable selon le jeu de données

La variable du numéro de cabine est celle, comme nous le voyons en table n°1, qui contient le plus de valeurs manquantes, l'âge est lui aussi assez incomplet puisque près de 20% des observations sont absentes de la base d'entraînement, et un peu plus dans la base test. Pour les variables '*Fare*' et '*Embarked*' on voit qu'il manque 1 ou 2 observations seulement pour les 2 jeux. Aussi, pour l'âge qui comporte un quart de données manquantes nous décidons d'imputer la médiane qui est de 28 ans pour le jeu train et 27 pour le jeu test. Nous faisons de même pour Fare et Embarked en imputant la médiane ou le mode. Enfin, nous laissons telle quelle la variable '*Cabin*' puisque nous procéderons à une transformation de celle-ci plus tard dans l'analyse.

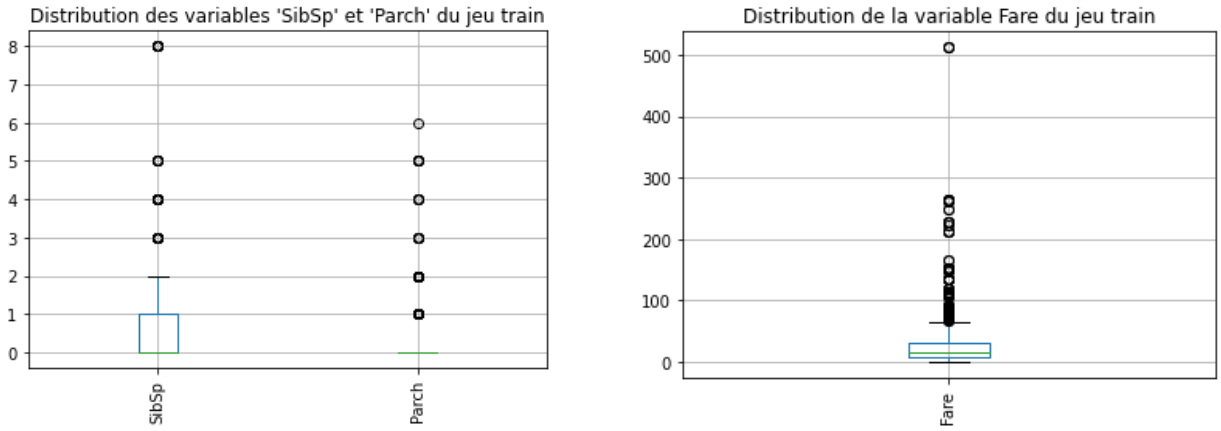


FIGURE 2 – Valeurs atypiques du jeu train

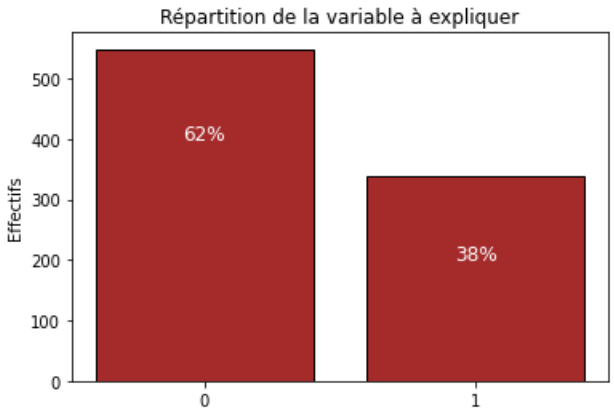
La figure n°2 nous montre quelques observations atypiques pour le nombre de frères et soeurs ou conjoints à bord, variable pour laquelle la valeur atteignait 8 tandis que la moyenne était de 0.52, et pour le prix du

billet qui dépassait 500£ alors que la moyenne était de 32£ par achat. Ce montant si important tient au fait que le prix indiqué corresponde à un achat qui peut inclure plusieurs billets à la fois, cependant le point atypique correspond au billet de Miss. Anna Ward qui est venue seule à bord et qui devait être fortunée puisqu'elle a payé 512,33£ sa place en première classe. Nous retirons ces quelques outliers et obtenons une base **train** aux dimensions (879x11), il y a une colonne de moins qu'à l'import puisque nous avons passé l'identifiant du passager comme indice des 2 bases.

Nos données sont désormais propres et avant de commencer les modélisations nous nous penchons sur chaque variable pour en extraire des informations qui pourront être utiles par la suite.

3 Statistiques descriptives

3.1 Variables quantitatives



La première des variables quantitatives est celle que nous cherchons à expliquer : la survie des passagers lors du naufrage du 15 avril 1912. Comme visible en figure ci-dessous, près de deux tiers des voyageurs sont décédés cette nuit là, seuls 38% des personnes ont survécu. On se demande alors si des caractéristiques communes sont identifiables au sein des groupes des survivants et des non survivants, pour cela nous regardons certaines distributions des variables selon la survie des passagers.

	Âge	**Prix**	**Parenté**	**Fratrie**
Décès	30.62	22.12	0.33	0.55
Survie	28.19	48.21	0.47	0.48

TABLE 2 – Caractéristiques quantitatives moyennes des passagers selon leur survie

En table n°2, en figure n°3 ainsi qu'en annexe n°1 nous voyons la distribution des variables explicatives quantitatives selon la survie des passagers ; il apparaît alors que les personnes les plus susceptibles de survivre sont celles qui ont payé plus cher leur billet, qui sont très jeunes (enfants) et qui sont venues avec un ou deux proches. On voit en effet qu'en moyenne les survivants avaient payé plus de deux fois plus cher leur place à

bord du bateau que les décédés, aussi, parmi les personnes qui ont payé entre 0 et 10/15£ leur place on voit qu'il y a plus de 3 fois plus de morts que de survivants. L'histogramme de distribution de l'âge nous informe que la seule classe d'âge où l'on retrouve plus de survivants que de morts est celle des très jeunes enfants dont l'âge est inférieur à 8 ans, mais pour tous les autres il y a plus de victimes que de survivants. Nous décidons alors de créer une variable binaire prenant la valeur *'true'* si la personne est âgée de moins de 8 ans et *'false'* sinon, puisque le facteur de jeunesse semble beaucoup jouer dans l'explication de la survie au naufrage. Dans les catégories de cette nouvelle variable "**is_child**" on retrouve 50 personnes de moins de 8 ans (soit 0.9% de la base train) et 829 de plus de 8 ans.

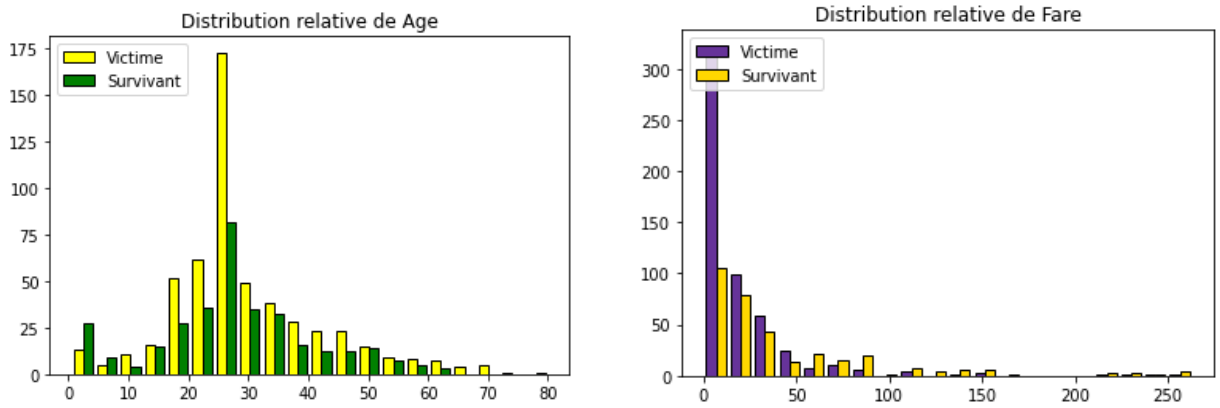


FIGURE 3 – Survie des passagers selon leur âge et le prix de leur billet

Aussi, pour finir cette analyse des variables quantitatives nous regardons la matrice de corrélation qui indique la causalité qu'il existe entre ces variables ; si le coefficient qui représente cette relation est supérieur à 0.5 alors nous considérons qu'elle est trop forte et serons contraints d'exclure une des 2 variables concernées lors des estimations (du moins pour l'application de modèles **économétriques**). D'après la figure n°4, il apparaît cependant que les coefficients n'excèdent pas 0.4, donc nous pourrons inclure toutes les variables quantitatives dans les modèles. On remarque tout de même que la corrélation la plus forte concerne le nombre de proches à bord, ce qui pouvait facilement s'anticiper.

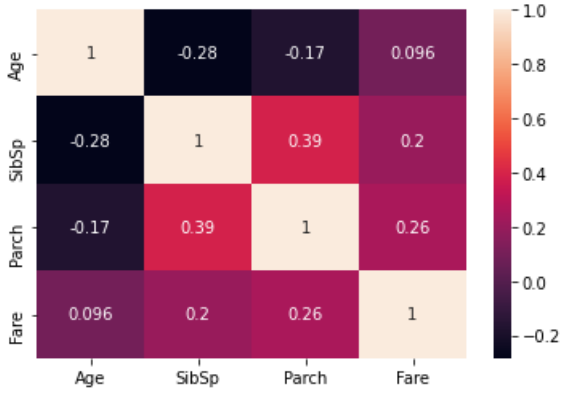


FIGURE 4 – Corrélations entre les variables explicatives quantitatives

3.2 Variables qualitatives

		Décès	Survie	Total
Classe :	1ère classe	80 (38%)	131 (62%)	211 (100%)
	2è classe	87 (47%)	97 (53%)	184 (100%)
	3è classe	365 (75%)	119 (25%)	484 (100%)
Sexe :	homme	464 (81%)	107 (19%)	571 (100%)
	femme	78 (25%)	230 (75%)	308 (100%)
Port :	Southampton	420 (66%)	217 (33%)	637 (100%)
	Cherbourg	90 (55%)	75 (45%)	165 (100%)
	Queenstown	47 (61%)	30 (39%)	77 (100%)
	Total	542 (62%)	337 (38%)	879 (100%)

TABLE 3 – Répartition des passagers dans les modalités selon leur survie

La table 3 nous informe de la répartition des passagers de la base *train* dans les différentes modalités des variables qualitatives, selon leur survie ou non. On voit d'emblée que les passagers en 3è classe étaient 2.3 fois plus nombreux que ceux en première et seconde classes, il y a près de 2 fois plus d'hommes que de femmes à bord (si l'échantillon est représentatif des vrais effectifs lors du voyage de 1912) et 72% des passagers ont embarqué au port de départ c'est-à-dire Southampton.

Concernant la variable à expliquer, on constate que le groupe de personnes avec le taux de survie le plus élevé est celui des femmes avec seulement 25% de décès, a contrario 81% des hommes de l'échantillon *train* ont perdu la vie. On note aussi que la classe qui a le taux de survie le plus important est la première classe : 62% des passagers survivent contre seulement 25% de la troisième classe. Le constat est bien triste mais il est réel ; le Titanic transportait à son bord les plus grandes fortunes de l'époque qui ont été sauvées en priorité, mais la survie de ce groupe peut aussi être liée à la logistique. Effectivement, le paquebot se décomposait en 10 ponts dont 7 pour les cabines allant ainsi de A à G. Or, comme visible en annexe n°2 les personnes les plus aisées voyageaient sur les ponts A, B et C, tandis que les 3è classes étaient sur les ponts E, F et G - c'est à dire les plus bas. Sachant que les canots de sauvetage, s'ils étaient remplis au maximum (ce qui n'était déjà pas le cas), pouvaient contenir 1178 personnes sur 2200 voyageurs et qu'ils étaient placés sur les ponts supérieurs, les passagers de 1ère classe y ont eu accès les premiers.

Enfin, on constate que les personnes ayant embarqué à Cherbourg sont celles qui ont le plus survécu avec un taux de 45% contre 33% pour les passagers de Southampton et 39% pour ceux de Queenstown qui sont les moins nombreux. Il est cependant difficile d'identifier une tendance à ce niveau puisque les effectifs diffèrent grandement entre le port de départ et les escales dans les villes de Nouvelle-Zélande et de France.

Cette partie de statistiques descriptives sur notre jeu de données d'entraînement nous aura permis de faire de nombreux constats et d'émettre des premières hypothèses quant aux facteurs influençant la survie des passagers. Dans une dernière sous-section nous allons créer de nouvelles variables qui nous semblent pertinentes pour les modélisations que nous commencerons en partie suivante.

3.3 Création de variables

Nous avons pu observer la distribution et la répartition des différentes variables explicatives de notre base. Il y en a pourtant 2 sur lesquelles nous ne nous sommes pas encore penchées puisqu'elles demandent une attention particulière, il s'agit du numéro de cabine et du nom du passager. Ces dernières ne sont pas traitables telles quelles (ni quantitatives ni composées de modalités), mais contiennent tout de même des informations qui nous seront utiles pour la suite de l'analyse.

À partir de la variable "**Name**" nous avons construit la variable '**title**' qui reprend uniquement l'information sur le titre de la personne mais ne prend pas en compte ses nom et prénom. En effet, le titre de chaque voyageur peut être un complément d'information sur la place de cette personne dans la société, complétant ainsi les variables 'Pclass' et 'Sex'. Parmi les titres les plus présents dans la base on retrouve :

- **Mr** avec 512 personnes soit 90% des hommes
- **Miss** avec 177 personnes soit 57% des femmes (il y avait donc beaucoup d'enfants chez les filles)
- **Mrs** avec 124 personnes soit 40%
- **Master** avec 39 personnes soit 7%

Comme nous l'avons expliqué précédemment, la variable '**Cabin**', bien que très incomplète, contient le numéro de la cabine avec d'abord une lettre allant de A à G correspondant à l'un des 7 ponts passagers, puis le numéro de la chambre. Pour les 679 observations pour lesquelles il nous manque cette information, nous attribuons la valeur *-1*, pour les autres nous ne gardons que la lettre correspondant au pont et obtenons les effectifs suivants :

- **A** : 15 passagers
- **B** : 43 passagers
- **C** : 59 passagers
- **D** : 33 passagers
- **E** : 32 passagers
- **F** : 13 passagers
- **G** : 4 passagers

Nous obtenons ainsi 2 nouvelles variables qui nous serviront pour les modélisations, par conséquent nous supprimons celles dont elles sont issues à savoir 'Cabin' et 'Name'. Nous avons évidemment réalisé ces transformations à la fois sur le jeu d'apprentissage et le jeu test qui doivent impérativement correspondre pour l'application de mêmes modèles. Dans la prochaine partie nous estimerons les modèles après avoir préparé les données pour.

4 Modélisations

4.1 Préparation des données

La préparation des bases aux modélisations passe par la standardisation des variables quantitatives et la catégorisation des variables qualitatives, avant de vérifier pour chacune d'elle le format grâce à la commande *'dtypes'*.

La standardisation des variables numériques est un prérequis à l'application de nombreux modèles ; la trop grande différence dans les échelles des variables peut altérer les résultats ou cacher une information pourtant intéressante dans les données. Nous corrigeons cela par la **normalisation** des 4 variables quantitatives : Age, Fare, SibSp et Parch, c'est-à-dire en soustrayant la moyenne et en divisant par l'écart-type. Les valeurs des variables standardisées s'étendent alors de -2 à 7 tandis qu'elles allaient jusqu'à 263 avant transformation.

Pour que les variables qualitatives soient prises en compte et traitées dans les modélisations, nous devons les passer en catégorie puis créer des "**dummy variables**". Au lieu d'avoir 3 modalités en une variable, passer par les *dummies* revient à créer autant de variables binaires différentes qu'il y a de modalité, en en gardant une de côté qui sert alors de référence. Nous appliquons cette transformation nécessaire pour l'application des modèles, grâce à la fonction "*get_dummies()*" de la librairie **pandas**. Nous passons alors de 10 à 35 variables explicatives dans le jeu "train".

Enfin, après avoir converti les variables au bon format et divisé les variables explicatives de Y (Survived), nous passons aux estimations en commençant par la régression logistique.

4.2 Estimations

Pour chacun des modèles que nous estimerons sur la base d'entraînement, nous évaluerons sa qualité grâce au *score* qui donne le taux de bonnes prédictions. Nous le regarderons à 2 niveaux : lorsque le modèle est appliqué sur les données sur lesquelles il s'est construit (train) et lorsqu'il est appliqué sur des données inconnues. Nous utiliserons pour cela la méthode de ré-échantillonage de **cross validation** qui divise aléatoirement les données d'apprentissage en k groupes de tailles égales. Le modèle se forme alors sur les k-1 groupes et s'entraîne sur le dernier - et ainsi de suite k fois. Quand les modélisations sont terminées, l'estimation de la CV correspond à la moyenne des k erreurs de test (nous choisissons k=3 dans notre cas). En outre, nous obtiendrons 2 taux de bien classés : sur des données connues et des données inconnues, le modèle final que nous retiendrons comme étant le meilleur sera celui qui maximise ces scores.

Pour l'ensemble des modèles que nous appliquons nous trouvons la fonction dans la librairie "**sklearn**". Dans un premier temps nous construisons une **régression logistique** faisant partie des modèles binaires qui s'appliquent lorsque l'on analyse une variable qualitative dite ' binaire', dans notre cas nous cherchons à expliquer les probabilités de survivre (1) ou de périr (0) dans le naufrage du Titanic. Le score de ce modèle sur la base train est de 82.58%, ce qui signifie qu'à partir des informations contenues dans la base d'entraînement, la régression arrive à prédire correctement le sort de 8 personnes sur 10. Avec 3 groupes en CV, nous obtenons les scores suivants : 79.18, 80.42 et 82.75% pour une moyenne de 80.78%. L'écart de bonnes prédictions entre

l'échantillon train et en validation croisée est inférieur à 2% ce qui montre qu'il n'y a pas de problème de sur-apprentissage, c'est-à-dire où le modèle s'ajuste trop aux données d'entraînement et est moins performant sur des données inconnues. Outre cet aspect technique important, la régression logistique offre la possibilité de connaître l'effet de chaque variable sur le phénomène à expliquer (par le calcul d'effets marginaux et odd ratios) et permet donc pleinement de répondre à la problématique de la compétition. Avant de conclure sur les facteurs influençant le fait de survivre ou non, regardons la table suivante qui reprend les informations liées aux différentes modélisations :

	Scores	**Train**	**CV**
	Reg. logistique	82.58%	80.78%
	CART	98.3%	75.00%
paramétrage	CART	77.75%	75.03%
	Forêt	96.73%	78.82%
paramétrage	Forêt	82.47%	79.99%
	Bagging	93.34%	80.97%
	Boosting	98.31%	79.65%
optimisation	Boosting	/	82.35%

TABLE 4 – Résultats des modélisations

Comme visible en tableau n°4 nous avons appliqué des modèles d'arbres décisionnels, de forêts aléatoires, de bagging et de gradient boosting. Pour chacun de ces premiers modèles de machine learning, nous voyons que le taux de bien prédit est très différent lorsque le modèle est appliqué sur des données connues et inconnues, ce qui met en avant le sur-apprentissage créé par de tels modèles. Par exemple l'arbre de décision (qui a la particularité d'être sujet au sur-apprentissage), prédit correctement le sort de 98 passagers sur 100 puisqu'il a pris en compte tout le 'bruit' associé aux observations du jeu **train**, en revanche il ne prédit correctement le sort que de 75 passagers sur 100 lorsque les informations sur ces derniers sont **nouvelles**.

Pour pallier à cela, nous avons précisé les paramètres des modèles de machine learning qui étaient sujets à l'overfitting, de manière à les rendre plus robustes notamment pour les appliquer sur le jeu test. Pour la forêt aléatoire nous avons par exemple précisé la profondeur des arbres à 7, le nombre d'estimateurs à 50 et le nombre minimum d'échantillons pour chaque fractionnement à 0.03. Grâce au paramétrage des différents modèles, nous obtenons des scores train et CV beaucoup plus proches qui traduisent des modèles robustes et pertinents. Pour l'arbre de décision, l'écart passe ainsi de 23 à 2 points de pourcentage.

Aussi, pour le modèle **gradient boosting** nous avons procédé autrement ; nous avons fait une recherche des meilleurs hyperparamètres grâce à la recherche en grille appelée **'grid search'** dont le but est de réduire l'erreur de prédiction tout en ne complexifiant pas trop le modèle pour qu'il reste robuste. Cette méthode explore les valeurs des hyper-paramètres proposées, et s'arrête lorsqu'elle a trouvé la combinaison qui minimise l'erreur de prédiction. Dans notre cas nous obtenons les paramètres '*learning_rate*' : 0.05, '*max_depth*' : 5 et '*n_estimators*' : 150, censés répondre au problème de sur-apprentissage. Nous obtenons alors un taux de bien classé de 82.35% par cross validation, qui fait de cette méthode de ML notre meilleur modèle.

4.3 Modèle final

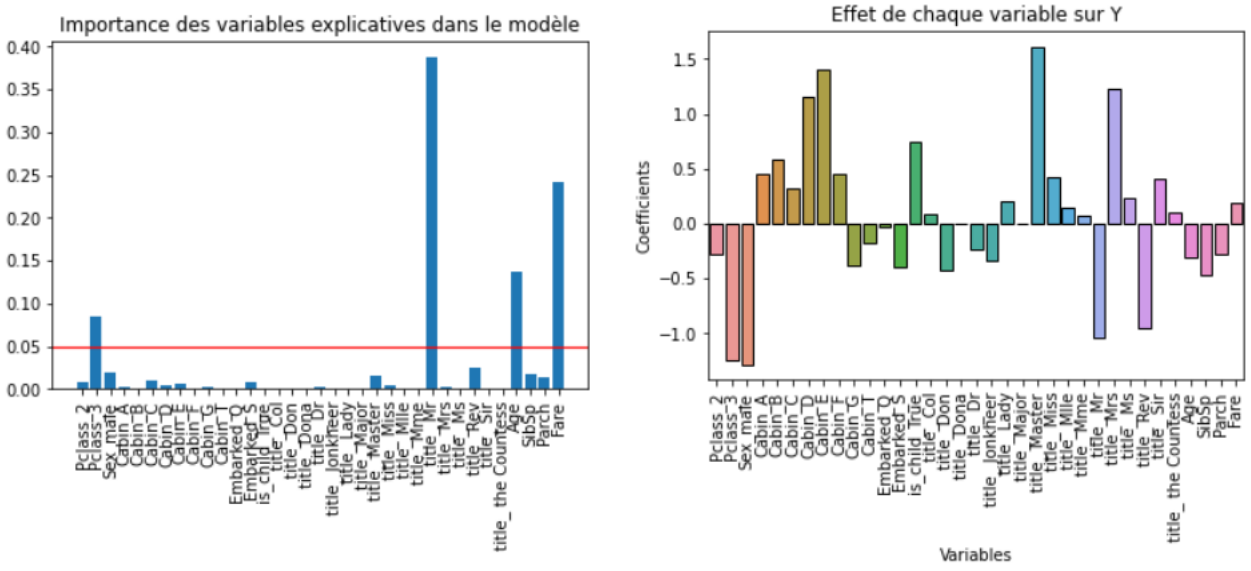


FIGURE 5 – Importance et effets des variables sur la survie des passagers

Le modèle que nous retenons finalement est donc le gradient boosting avec optimisation des hyper-paramètres pour lequel nous obtenons le *score* le plus élevé par validation croisée. La commande "*.feature_importances_*" sur python nous offre un aperçu des variables importantes pour ce modèle. En outre, comme visible en figure n°5 les variables les plus importantes du modèle gradient boosting sont la classe lors du voyage et plus particulièrement le fait d'être en troisième classe, le fait d'être un homme, l'âge et le prix du billet. Ce sont les 4 significativement importantes dans l'explication de la survie, au seuil de risque de 5% signifié par la ligne rouge. À présent, nous trouvions intéressant de quantifier l'effet (positif ou négatif) de chacune d'elle sur Y, nous avons pour cela récupéré les coefficients de la régression logistique qui est la seule à pouvoir quantifier les effets, nous l'avons fait grâce à la commande "*lr.coef_*", puis nous avons mis sur un graphique ces coefficients. Si l'on reprend les variables significatives du modèle boosting, on voit que les passagers de 3è classe avaient 1.24 fois moins de chances de survivre par rapport aux passagers de 1ère classe. Les hommes ne portant aucun titre (les 'Ms') avaient 1.04 fois moins de probabilité de s'en sortir, aussi, plus un passager est âgé et plus ses chances de survie diminuent (nous supposons ici des effets de seuils sur cette variable, il pourrait donc être intéressant de construire des modèles GAM pour compléter notre analyse). Enfin, plus une personne a payé cher son billet plus elle a de chances de survivre - ce qui rejoint les conclusions émises à partir de la variable 'Pclass'.

Si nous regardons les effets des autres variables bien que non significatives dans le modèle final, nous voyons qu'être sur les ponts F et G (donc les troisièmes classes cf. annexe n°2) diminue les probabilités de survie, tandis qu'être sur les ponts D et E les augmente grandement : cela peut être lié au fait que le chargement des rescapés dans les canots de sauvetage différait entre les côtés bâbord et tribord. Les ordres de l'un étaient "Les femmes et les enfants d'abord", tandis que l'autre (l'officier Murdoch à tribord) complétait par des hommes, remplissant ainsi davantage les canots. On note aussi que les hommes ayant le titre de **'Master'** avaient 1.61 fois plus de chance de survivre ; ce diplôme de 3è cycle traduit un niveau d'études élevé et rejoint donc la richesse du passager ainsi que sa classe (son pont) dans le paquebot.

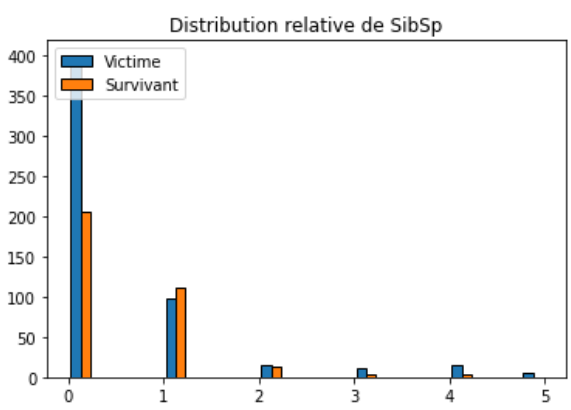
5 Conclusion

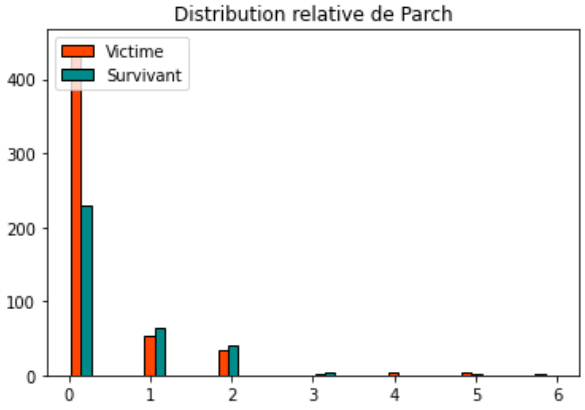
Pour conclure nous pouvons dire que le naufrage du 15 avril 1912 a été un drame maritime qui a causé la mort de 1500 personnes qui n'avaient pas les mêmes chances de survie à partir du moment où le Titanic a percuté l'iceberg. Les passagers les plus enclins à survivre étaient donc les enfants, les femmes, et les personnes "importantes" (par leur titre) qui pouvaient ainsi payer des billets plus cher. En effet, nous avons souligné le fait que les ponts des premières classes étaient situés plus haut dans le bateau et donc plus proches des canots de sauvetage qui nous le rappelons, pouvaient contenir seulement 1 178 personnes soit un peu plus de la moitié des passagers. Ajoutons à cela 2 choses, l'une que ces canots en plus d'être en nombre insuffisant n'ont été remplis qu'au deux tiers en moyenne, l'autre que les personnels étant beaucoup moins nombreux pour les 2è et 3è classes, ils ont mis plus de temps à prévenir les passagers accentuant alors davantage l'aspect inégalitaire face à l'accessibilité des canots de sauvetage.

D'un point de vue technique, nous avons créé 2 variables à partir des existantes : le titre du passager à partir de son nom complet, et le numéro du pont à partit de la variable 'Cabin'. Elles se sont révélées pertinentes puisque parmi les 4 modalités de variables retenues par le modèle final de gradient boosting, une correspondait au titre "Mr" créé par nous-mêmes. Finalement, lorsque nous appliquons le modèle final au jeu de données test (dont les observations sont donc inconnues au modèle) et que nous soumettons sous kaggle, nous obtenons un score de 0.7488 ce qui est satisfaisant compte tenu de nos recherches. Une possible amélioration que nous avons soulignée précédemment, consisterait à appliquer des fonctions sur les variables quantitatives pour préciser leur relation avec Y, et ainsi améliorer la qualité de nos prévisions.

6 Annexes

Annexe n°1 : Histogramme de distribution des variables 'SibSp' et 'Parch'.





Annexe n°2 : Décomposition des ponts du Titanic.